NPS55-81-001

NAVAL POSTGRADUATE SCHOOL

Monterey, California

A 3-D DATA SMOOTHING ALGORITHM

by

J. B. Tysver

January 1981

Approved for public release; distribution unlimited.

Prepared for: Research and Engineering Department
Naval Undersea Warfare Engineering Station
Keyport, Washington 93845

NAVAL POSTGRADUATE SCHOOL
MONTEREY, CALIFORNIA

Rear Admiral J. J. Ekelund
Superintendent

D. A. Schrady
Acting Provost

The work herein was supported in part by funds provided by the Naval Undersea Warfare Engineering Station, Keyport, Washington.

This report was prepared by:

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NPS55-81-001	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* A 3-D Data Smoothing Algorithm		5. TYPE OF REPORT & PERIOD COVERED Technical 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR*(s)* J. Bryce Tysver		8. CONTRACT OR GRANT NUMBER*(s)*
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Postgraduate School Monterey, CA 93940		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS N0025381WR70012
11. CONTROLLING OFFICE NAME AND ADDRESS Research and Engineering Department Naval Undersea Warfare Engrg. Station Keyport, WA 98345		12. REPORT DATE January 1981 13. NUMBER OF PAGES 36
14. MONITORING AGENCY NAME & ADDRESS*(if different from Controlling Office)*		15. SECURITY CLASS. *(of this report)* Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Noise
Missing points
Outliers
Sequential differences
Least squares regression
Polynomials

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

The Algorithm presented in this report is designed for use in smoothing 3-D data at NUWES. It uses fourth-order sequential differences to screen the data for outliers and then a special form of least-squares smoothing is performed to select an appropriate low-order polynomial and fit it to seven-point data segments.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

A 3-D DATA SMOOTHING ALGORITHM

by

J. B. Tysver

Naval Postgraduate School
Monterey, CA 93940

January 1981

I. INTRODUCTION

The algorithm proposed in this report for smoothing 3-D data of NUWES uses sequential differences to screen the data for "wild" data values (outliers) and a 7-point least-squares procedure to perform the data smoothing.

It is assumed that the actual relationship between each coordinate (x,y,z) of the path of the vehicle being tracked and time can be expressed as a low order polynomial $P_k(t)$ where k is the order (degree) of the polynomial. An observed value X_i can then, for example, be expressed as

$$x_i = P_k(t_i) + n_i$$

where n_i is the noise component resulting from inaccuracy in measuring the x-component of the position of the vehicle at time t_i. Examination of 3-D data from a torpedo run at NUWES, supported by consideration of the relationship between vehicular maneuvering capabilities and data rate, suggests that polynomials of no higher order than $K = 3$ (cubics) need be considered for 7 consecutive point data segments. The selection of the appropriate order of polynomial and the smoothing are incorporated in STEP 4 of the algorithm and discussed in Section III D.

Performance of least-squares smoothing can be standardized and computational requirements considerably reduced when 7-point data segments contain no missing values. This necessitates provision of temporary values to fill in gaps in the

data. Such temporary values are also useful in the screening process to identify outliers. Provision of temporary values for missing data values is treated in STEP 1 of the algorithm and discussed in Section III A.

When a component of a vehicular path is a polynomial of order $k \leq 3$, then the fourth order sequential differences of the observational data will consist exclusively of the noise components in the data, (Ref. 2). Since outliers can be attributed to noise it would appear reasonable to use fourth order sequential differences to screen for them. This is incorporated in STEP 2 and discussed in Section III B. This step calls for replacement of outliers by temporary values as in the treatment of missing values in STEP 1.

The temporary values to fill in missing points and replace outliers were selected to be consistent with the closest observed data value before and the closest after the point in question. This could, and actually does in some cases, produce temporary values that are inconsistent with the polynomial that best fits the observed values in the data segment. It is proposed in STEP 3 that smoothing be performed as in STEP 4 for 7-point data segments centered at the temporary values and that these smoothed values be treated as observed values in smoothing at observed data values. This is a direct attempt to minimize the effects of temporary values (either missing points or outliers) on the results of data smoothing. This is discussed in Section III C.

The presence of two or more missing points and/or outliers (i.e., temporary values) within the same 7-point data segment, or, more specifically, within three data points of each other, requires special treatment. Some initial efforts in this direction are included in STEP 5 and discussed in Section III E. The general motivation for this step is the derivation of the maximum information on a vehicular path that can be obtained from the data even though that information is degraded by missing points and outliers. The possibility of providing some measure of the "quality" of the smoothing at each data point for the guidance of potential users is discussed in Section III F.

II. A DATA SMOOTHING ALGORITH

The following Algorithm using fourth order sequential differences and least-squares smoothing of 7-point data segments is proposed for smoothing of 3-D data at NUWES.

STEP 1 Preliminary treatment of missing data points.

(a) A missing data value at time t_i where the adjacent observed values x_{i-1} and x_{i+1} are available is assigned the temporary value

$$x_i^* = (x_{i-1} + x_{i+1})/2 \ .$$

(b) Adjacent missing values at times t_i and t_{i+1}, where the values x_{i-1} and x_{i+2} are observed, are assigned the temporary values

$$x_i^* = (2x_{i-1} + x_{i+2})/3 \quad \text{and}$$

$$x_{i+1}^* = (x_{i-1} + 2x_{i+2})/3 \ .$$

(c) Occurrence of more than two consecutive missing values are designated for Special Treatment in Step 5.

STEP 2 Identify and temporarily replace Outlier data points.

(a) Calculate fourth order sequential differences d_{4i}'s . (Ref. 2)

(b) Any value d_{4i} which exceeds a specified level d in magnitude is a potential outlier. The signature in fourth order differences of an outlier is a large value d_{4i} with somewhat smaller values d_{4i-1} and d_{4i+1}

of opposite sign adjacent. (These adjacent values may, or may not, be outliers even though they exceed d in magnitude.) Designate x_i as an outlier and treat it as a missing value as in Step 1.

(c) Repeat Step 2(a) and Step 2(b) in the neighborhood of x_i .

STEP 3 Determine smoothed values to replace temporary values.

(The 7-point least-squares smoothing procedure of STEP 4 is used to determine smoothed values to replace the temporary values obtained in STEP 1 and STEP 2.)

(a) Any temporary data value x_i^* for which the 7-point data segment $(x_{i-3}, x_{i-2}, x_{i-1}, x_i^*, x_{i+1}, x_{i+2}, x_{i+3})$ contains no other temporary values is treated as follows.

(i) Apply the 7-point least-squares smoothing procedure of STEP 5 to determine a smoothed value x_i' .

(ii) If the difference $x_i^* - x_i'$ is less than a prescribed value a (Section III C) then x_i' is accepted as the smoothed value of x at time t_i . If $|x_i^* - x_i'| > a$, then replace x_i^* by x_i' and repeat step (i) above. This process is repeated until a difference $|x_i^* - x_i'|$ less than a is achieved.

(b) If two temporary data values x_i^* and x_j^* are separated by less than three observations (j = i+1, i+2, or i+3) and a 7-point data segment centered at x_i^* or x_j^* (or, preferably, centered between x_i^* and x_j^*) contains no other temporary values, then the least-squares smoothing procedure of STEP 5 is applied to that data segment to determine smoothed values x_i' and x_j' simultaneously as in (a) above.

(c) Any temporary value that cannot be smoothed by (a) or (b) above is designated for special treatment (STEP 5).

STEP 4 Smooth observed data values.

Each observed data value which is at the center of a 7-point data segment containing observed values or previously smoothed values (STEP 3) is smoothed using the 7-point least-squares smoothing procedure described in Section III D. Other observed data values are designated for special treatment.

STEP 5 Special treatment of designated data points.

The purpose of this step is to determine additional information, wherever possible on the vehicular path in the neighborhood of multiple outliers and/or missing points. (It should be recognized, and specifically indicated to potential users of the smoothed data that there is greater uncertainty in the actual vehicular path when STEP 5 is used for smoothing.) Two possibilities for accomplishing this are included here. Other possible treatments may be added as they are developed.

(a) Observed data values which do not satisfy the requirements for treatment in STEP 4 but are within a 7-point data segment containing only observed or previously smoothed data values are also to be smoothed as in STEP 4. In this case the center of the 7-point segment used for smoothing should be as close to the observed value to be smoothed as possible. (This procedure could also be used to predict values outside the 7-point segment with increasing uncertainty.)

(b) There is some possibility of extracting additional information on the vehicular path in the vicinity of multiple temporary values by relaxing the requirement in STEP 3 (b) on the location of the center of the 7-point segment used for smoothing with respect to the temporary values to be smoothed.

III DISCUSSION

A. MISSING DATA POINTS (STEP 1)

The role of STEP 1 is to provide temporary values for missing data points. These temporary values are used in two ways. First, they are of help in reducing gaps in fourth order sequential differences and hence permit a more complete examination of those differences in the search for potential outlier values. Second, they provide initial values for the 7-point smoothing procedure which, in its special form as described in Section III D, is only applicable to segments of seven sequential data values.

Perhaps the simplest way to provide temporary values for missing points is to assume that the actual relationship of x (for example) and t is linear and hence to use linear interpolation for missing values between adjoining observed values. For a missing value at time t_i when x_{i-1} and x_{i+1} are the adjacent observed values, the appropriate temporary value at time t_i then becomes

$$x_i^* = x_{i-1} + \frac{1}{2}(x_{i+1} - x_{i-1}) = (x_{i-1} + x_{i+1})/2 \ .$$

For adjacent missing values at times t_i and t_{i+1}, use of the observed values x_{i-1} and x_{i+2} yields the temporary values

$$x_i^* = x_{i-1} + \frac{1}{3}(x_{i+2} - x_{i-1}) = (2x_{i-1} + x_{i+2})/3 \quad \text{and}$$

$$x_{i+1}^* = x_{i-1} + \frac{2}{3}(x_{i+2} - x_{i-1}) = (x_{i-1} + 2x_{i+2})/3 \ .$$

For missing values at times t_{i-1}, t_i, and t_{i+1}, the appropriate temporary values are

$$x^*_{i-1} = x_{i-2} + \frac{1}{4}(x_{i+2} - x_{i-2}) = (3x_{i-2} + x_{i+2})/4 ,$$

$$x^*_i = x_{i-2} + \frac{2}{4}(x_{i+2} - x_{i-2}) = (x_{i-2} + x_{i+2})/2 , \quad \text{and}$$

$$x^*_{i+1} = x_{i-2} + \frac{3}{4}(x_{i+2} - x_{i-2}) = (x_{i-2} + 3x_{i+2})/4 .$$

Extension to longer sequences is not useful for either screening for outliers or for data smoothing.

B. OUTLIERS (STEP 2)

As indicated in Section I, it is assumed that the relationship between a component of the path of a vehicle being tracked can be represented by, for example, x as a polynomial function of t of no higher order than three. The fourth order sequential differences of the observed values, then, will be functions of the noise components in the observed values. It was shown in Reference 2 that the standard deviation of the fourth order sequential difference d_{4i} at time t_i is

$$\sigma_{4i} \doteq 8.367\sigma_N$$

where σ_N is the standard deviation in the noise component of each observed value.

In order to be useful this must be translated into a threshold value d with any observed value x_i for which d_{4i} exceeds d in magnitude being identified as an outlier. It

would appear reasonable to assume that d_{4i}, as a linear combination of noise components, is at least approximately normally distributed with zero mean and hence the probability that D_{4i}, as a random variable, will exceed $3\sigma_{4i}$ in magnitude is less than 0.01. The threshold value d can then be expressed as

$$d = 3\ \sigma_{4i} \doteq 25\ \sigma_N\ .$$

(Note that in a sequence of 100 observed values one legitimate observed value, on an average, will be incorrectly identified as an outlier.)

There remains the problem of specifying an appropriate value for the standard deviation σ_N of the observational noise. There are several possible approaches here, all of which rely basically on estimation of σ_N by the standard deviation(SE) of the residual errors after fitting data segments by polynomials using the least squares method. (Ref. 1) The following ways of estimating σ_N are possible:

(1) Historical Data - Fitting of polynomials to data segments from 3-D data on torpedo paths at NUWES yielded values $SE \doteq 2$ in many cases. (Ref. 1 and Table 1 to follow)

(2) Technical - Information should be available from instrumentation personnel on the capabilities of the position location system in use at NUWES. (This has not been explored.)

(3) Current Data - A possible substitute for historical data would be the use selected segments from the data to be smoothed. Trial fitting of these segments solely for the purpose of obtaining current values of SE could be used to estimate σ_N. (This approach has the advantage that it represents the current status of the position location system. It has the disadvantage that it can be seriously contaminated by outlier points which cause large values of SE.) (Isolated large values of SE could be considered as an indication of the presence of potential outliers. This could be followed up to determine whether the large value of SE was caused by a single large residual error. This could be incorporated into the algorithm to provide a second screening for outliers performed whenever an unusually large value of SE occurs in the data smoothing (STEP 3 and STEP 4)).

3-D data from a torpedo path on a run at NUWES was used to assist in the development of the algorithm. In this data every eighth data point was missing due to the data collection procedure. Trial sample variances were calculated for 30 of the data segments between these missing points. (Missing points in the segments were treated as in STEP 1 and screening for outliers was not performed in these calculations.) Using the least-squares procedure described in Section III D, sample variances SEK of

residual errors were calculated for linear (K=1), quadratic (K=2) and cubic (K=3) polynomials. These values and the value of K for which SEK is smallest are presented in Table 1. The last column of the table indicates data segments containing observed values identified as outliers in STEP 2 using $\sigma_N \doteq 2$ so that

$$d \doteq 25\ \sigma_N \doteq 50 .$$

(The average value of the minimum SEK's for each segment, omitting the segments with indicated outliers, is 2.10.

(4) Position Dependent Thresholds - It should be recognized that the magnitude of σ_N can vary with such factors as the location of the vehicle being tracked with respect to the position location of the array collecting the data and the state of the transmission medium in different parts of the test area. It is then conceivable that different values of σ_N could be used to specify different values of d depending on the part of the test area where the segments of the vehicular path occurred.

Performance of STEP 1 and STEP 4 for the data used to produce Table 1 yielded the sequence shown below for the data segment centered at time $t_i = 2157$.

TABLE 1

SAMPLE STANDARD DEVIATIONS AND OUTLIERS
(X-COMPONENT)

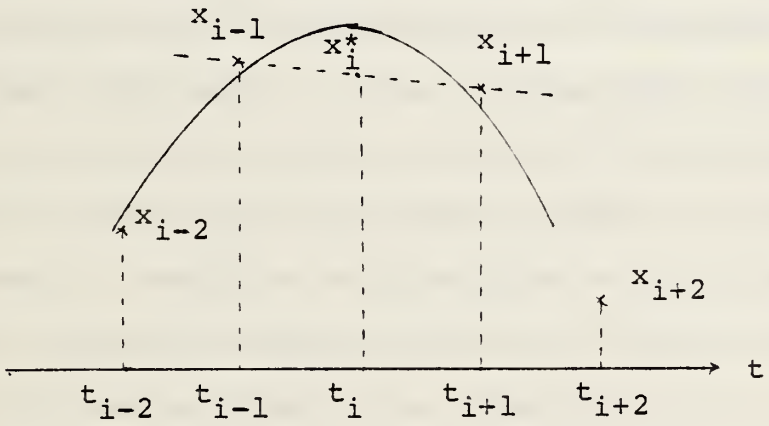
j (Segment)	t_j (Time)	SE1 (Linear)	SE2 (Quadr.)	SE3 (Cubic)	K (Min SEK)	Outlier Indicated
1	2092	21.61	7.46	8.40	2	X
2	2100	28.51	5.44	2.34	3	
3	2108	18.49	5.44	2.64	3	
4	2116	2.92	2.14	1.90	3	
5	2124	6.46	2.02	2.33	2	
6	2132	35.74	18.58	2.84	3	
7	2140	21.79	23.24	1.85	3	
8	2148	9.81	5.61	4.04	3	
9	2156	26.47	15.82	15.31	3	X
10	2164	7.74	2.69	3.08	2	
11	2172	19.80	2.99	3.33	2	
12	2180	15.74	5.17	2.64	3	
13	2188	1.42	0.81	0.90	2	
14	2196	3.04	1.21	1.02	3	
15	2204	30.96	6.71	7.72	2	X
16	2212	28.40	19.68	2.58	3	
17	2220	15.47	13.15	4.95	3	X
18	2228	24.59	2.99	2.44	3	
19	2236	3.80	3.49	3.73	2	
20	2244	24.01	1.62	0.52	3	
21	2252	11.90	3.91	3.45	3	
22	2260	1.36	1.47	1.22	3	
23	2268	2.25	2.47	2.13	3	
24	2276	34.78	18.96	2.23	3	
25	2284	60.60	14.96	8.87	3	X
26	2292	36.58	5.65	2.10	3	
27	2300	15.70	2.60	0.19	3	
28	2308	10.53	5.24	0.5	3	
29	2316	29.12	1.34	1.37	2	
30	2324	12.35	8.11	7.62	3	X

t_i	x_i	d_{4i}
2154	33,219.5	- 13.4
2155	33,212.5	23.1
2156	33,221.0	- 130.8
2157	33,273.5	212.2
2158	33,267.7	- 146.7
2159	33,313.5	21.9
2160	33,374.2*	29.8

As indicated in Reference 2, an isolated outlier will contaminate the fourth order differences of the adjacent points producing difference of substantial magnitudes of opposite sign. Using a threshold value of $d = 50$, the value of x_i at $t_i = 2157$ is strongly indicated as an outlier value. Replacing this x_i by x_i^* using STEP 1 and recalculating values for d_{4i} in the neighborhood of $t_i = 2157$ produced the modified results shown below.

t_i	x_i	d_{4i}
2154	-	
2155		
2156		- 14.2
2157	33,244.35*	37.3
2158		- 30.0
2159		
2160		

The possibility of a temporary value created in STEP 1 for a missing point being indicated as an outlier in STEP 2 is interesting although it requires no special treatment. As indicated in Section III A, a temporary value x_i^* is supplied for an isolated missing point by taking the average of the observed values on each side of it. As illustrated below, where the actual x component of the vehicular path is a quadratic (parabolic) the temporary value can differ substantially from $P_2(t_i) + n_i$ and produce a large value of d_{4i}.



Three of the 30 missing points in the data for the path presented in Table 1 were supplied temporary values using STEP 1 and which were identified in STEP 2 as outliers. One of these occurred at time $t = 2136$. Data for the 7-point segment centered at this time is presented below.

i	t_i	x_i	d_{4ix}	d_{4iy}	d_{4iz}
1	2133	33,637.7	- 23.4	- 12.0	26.5
2	2134	33,556.5	34.3	- 0.2	- 16.2
3	2135	33,486.6	- 88.2	- 19.3	0.7
4	2136	33,466.5*	108.7	24.0	1.9
5	2137	33,446.3	- 82.5	- 22.3	0.9
6	2138	33,485.1	4.8	14.5	- 6.9
7	2139	33,559.3	- 0.6	10.3	15.3

The outlier in the x-component indicated at time 2136 is not accompanied by outliers indicated in the y and z-components. This suggests that temporary values can be designated as outliers for one component (x) without that designation being necessary for the other components (y and z). Note also that although d_{4i} exceeds the threshold $d = 50$ in magnitude at times $t = 2135$ and $t = 2137$, the corresponding values of x_i should not automatically be designated as outliers. Supporting evidence that they are not outliers comes from Table 1 where these x_i's lie in data segments $j = 6$ and 7 respectively and the standard deviations for those segments SE_j are 2.84 and 1.85 which suggests that large residual errors are not present at either of these times.

On the other hand, the outlier in the x-component at the time $t = 2157$ and previously discussed is accompanied by the values $d_{4iy} = - 139.2$ and $d_{4iz} = - 193.8$ so that all 3 components have outliers indicated at $t = 2157$. This brings up the question as to whether a legitimate outlier (not a temporary value at a missing point) can occur in one component

only or that an outlier indicated in one component should require that the observed values of the other components are also suspect and should be treated as outliers whether their fourth order differences exceed d or not.

C. SMOOTHED VALUES FOR TEMPORARY VALUES (STEP 3)

The standardized computational format for smoothing seven consecutive data using the least-squares method is discussed in the next section (Section III D). It is proposed that this procedure be used repetitively on the 7-point data segment centered at the temporary value x_i^*. At each repitition, the temporary value x_i^* is replaced by the smoothed value. This iteration is continued until the residual noise $e_{ij} = x_i^* - x_i'$ is reduced to some acceptable level. (Theoretically, it could be repeated until $e_{ij} \doteq 0$). To illustrate this consider the example with an outlier at $t_i = 2157$. (Note: This is not at the center of the data segment examined in what follows but the shift of the temporary value from $i = 0$ to $i = 1$ was originally dictated by the fact that every eighth data point (values at $t_i = 2152$ and 2160) was missing. The results presented here involve use of the 7-point segment between times 2152 and 2160 and have not been recalculated.) The results of smoothing on the outlier value x_i and of three repititions of smoothing starting at the temporary value x_i^* (STEP 1) are shown below.

smoothing stage j	x_{ij}	SE_j	e_{ij}	x'_{ij}
outlier	33,273.5	15.3	19.42	33,254.1
1	33,244.35*	3.49	3.45	33,240.9
2	33,240.9	2.53	1.57	33,239.3
3	33,239.3	2.28	0.69	33,238.6

The smoothing procedure could have terminated after the second stage since the residual $e_{i2} = 1.57$ is less than $SE_2 = 2.53$ and hence well within the noise level of the other observations in the segment. One additional stage was used to produce the final smoothed value $x'_i = 33,238.6$ at time $t_i = 2157$. (Additional stages could reduce e_{ij} further but would have decreasing effect on SE_j.)

It is of interest to see the effect on successive differences (STEP 2) of replacing the outlier value $x_i = 33,273.5$ by the smoothed value 33,238.6. This is shown below.

i	pre d_{4i}	post d'_{4i}
- 3	- 2.6	
- 2	- 13.4	- 13.4
- 1	33.1	1.8
0	- 130.8	- 8.4
1	212.2	2.8
2	- 146.7	- 7.1
3	21.9	

The treatment for adjacent temporary values, whether resulting from missing points or outliers, is similar. Instead of determining smoothed values for one of them and then the

other or of alternating the smoothing stages between the two (and alternating data segments also), it would appear reasonable to smooth both simultaneously with the data segment centered on either one. (If one segment has additional temporary values but the other does not, the segment without additional missing points is to be used.) At each stage both temporary values are replaced by their smoothed values.

For two temporary values separated by one or more observed values the relative merits of simultaneous smoothing versus alternate smoothing stages centered first on one temporary value then on the other has not been examined. The widths of confidence intervals when the actual relationship is linear ($x(t) = P_1(t)$) increases with distance from the midpoint. (Ref. 1) This could suggest that simultaneous smoothing might still be appropriate if the temporary values are not too widely separated. In this procedure the 7-point data segment should have its center as close to the midpoint between the temporary values as possible. A modification of this selection may be required to avoid other temporary values, however. (This modification could be used in STEP 5 for the treatment of multiple temporary values.)

The presence of three temporary values in a data segment of seven consecutive points causes special difficulties. Iterated simultaneous smoothing of the three values could, theoretically, be carried to the limit in which a cubic equation involving four parameters is fitted exactly to the four observed values in the data segment. This results in all

residual errors, and hence SE, being equal to zero. The cubic is fitted not just to the path but to the noise as well. It would be possible to carry the smoothing only to the stage where the residual errors at the times of the temporary values are within the noise.

D. LEAST-SQUARES SMOOTHING (STEP 4)

The basic elements of the Least-Squares Method for smoothing two-dimensional data is described in Reference 1. For the purposes of the algorithm being developed, the method will be adapted to data segments of seven consecutive data values equally spaced in time. The procedure will be presented for the x-component only with the fitted function being of the form

$$\hat{x}(t) = b_0 + b_1 t + \ldots + b_k t^k ,$$

using 7 pairs of data of the form (t_i, x_i).

The computations involved in fitting the function $x(t)$ to the data is facilitated by making the following time translations of the data.

t_i	t_{i-3}	t_{i-2}	t_{i-1}	t_i	t_{i+1}	t_{i+2}	t_{i+3}
i	-3	-2	-1	0	+1	2	3
x_i	x_{-3}	x_{-2}	x_{-1}	x_0	x_1	x_2	x_3

This transformation serves two purposes. First, it reduces the magnitudes of the numbers to be computed and hence reduces

inaccuracies caused by computer round-off. To illustrate this, consider the dat segment including the time $t_i = 2157$. Fitting a polynomial of degree three to this data segment involves the term $SS3 = \sum t_i^6$ which is of the order of magnitude of 10^{20} and will be severely reduced in significant digits by any computer. Translation from t_i to i changes this term to $SS3 = \sum_{-3}^{+3} i^6 = 1588$. The important information in $SS3 = \sum_{1}^{7} t_i^6$ is contained in the last 8 digits which are lost in round-off rather than the first 8 digits which are retained by a computer using 8 decimal places.

The second, and equally important, purpose of the translation in time is that the portions of the regression calculations relating exclusively to the t_i's (the i's in the translated times) are common to all 7-point data segments and can be precalculated and introduced into the computations as constants.

A translation of the x_i's of the form

$$x_i = x_i^0 - \bar{x}$$

where x_i^0 indicates the observed value of x at time i and

$$\bar{x} = \frac{1}{7} \sum_{-3}^{3} x_i^0$$

will also help reduce the effects of computer round-off for such terms as $SSX = \sum x_i^2$.

Note that these translations in t_i and x_i produce the relationships

$$\sum_{-3}^{3} i = \sum_{-3}^{3} i^3 = \sum_{-3}^{3} i^5 = \sum_{-3}^{3} x_i = 0$$

and these terms drop from the formulas providing additional simplification in the computations. It will be assumed that the use of the data smoothing portions of the algorithm is preceded by these translations of x and t.

The information from the data segment essential to the smoothing procedure includes mean value $\bar{x}$ and the following statistics:

$$S1X = \sum i x_i,\ S2X = \sum i^2 x_i,\ S3X = \sum i^3 x_i, \text{ and } SSX = \sum x_i^2 .$$

(All of the summations are over i = -3 to i = +3.) Intermediate values to be calculated from the above and which are needed are:

$$SSX.1 = [28(SSX) - (S1X)^2]/28,$$
$$S3X.1 = (S3X) - 7(S1X),$$
$$SSX.2 = [84(SSX.1) - (S2X)^2]/84, \text{ and}$$
$$SSX.3 = [216(SSX.2) - (S3X.1)^2]/216 .$$

One of the advantages of the least-squares method of fitting polynomials to data is that the standard deviations of the residual errors can be calculated for any order polynomial (K = 1,2,3 for this algorithm) before any of the polynomials are fitted and hence before smoothed values and residual errors are determined. Thus

$$SE1 = \sqrt{\frac{1}{n-2} \sum e_{1i}^2} = \sqrt{(SSX.1)/5} \ ,$$

$$SE2 = \sqrt{\frac{1}{n-3} \sum e_{2i}^2} = \sqrt{(SSX.2)/4} \ , \quad \text{and}$$

$$SE3 = \sqrt{\frac{1}{n-4} \sum e_{3i}^2} = \sqrt{(SSX.3)/3} \ .$$

These standard deviations of residual errors provide a basis for selection of the appropriate order of polynomial with the order selected having the smallest value of SEK. (This has already been illustrated in Table 1.)

As a consequence of the nature of the least-squares method of fitting polynomials, the sums of squares of the residuals must satisfy the inequalities

$$\sum e_{1i}^2 \geq \sum e_{2i}^2 \geq \sum e_{3i}^2 \geq - - - -$$

The possibility of a second order polynomial, for example, producing a smaller value SE2 than a third order polynomial is due to the divisor n-(k+1) which accounts for the fact that a polynomial of order K has K+1 coefficients. (This divisor is called "degrees of freedom"

$$\nu = n-(k+1) \ .$$

It will be examined again in Section III F.)

For a 7-point data sequence (n=7), a third order polynomial will be considered to give a "better" fit than a second order polynomial (SE3 < SE2) only if

$$\sum e_{3i}^2 < \frac{3}{4} \sum e_{2i}^2 \ .$$

This contradicts our belief that a smaller value for $\sum e_{ki}^2$ indicates a closer fit of the polynomial. The statistic SEK is selected as the criterion here because its square is an unbiased estimate of the noise variance σ_N^2 and hence SEK is used in establishing confidence intervals for actual components of the vehicular path.

Note that for 9-point data segments SE3 will be less than SE2 whenever

$$\sum e_{3i}^2 < \frac{n-4}{n-3} \sum e_{2i}^2 = \frac{5}{6} \sum e_{2i}^2 .$$

A smaller decrease in $\sum e_{3i}^2$ will lead to selection $k = 3$ than when $n = 7$. (This is one reason for statistician's insatiable demand for larger samples.)

The coefficients for the selected order polynomial can now be calculated as follows:

Linear	$x'(t) = b_{10}+b_{11}i$	$b_{11} = (S1X)/28$
	$= P_1'(t)$	$b_{10} = 0$
Quadrabic	$x'(t) = b_{20}+b_{21}t+b_{22}t^2$	$b_{22} = (S2X)/84$
	$= P_2'(t)$	$b_{21} = (S1X)/28$
		$b_{20} = -4b_{22}$
Cubic	$x'(t) = b_{30}+b_{31}t+b_{32}t^2+b_{33}t^3$	$b_{33} = (S3X.1)/216$
	$= P_3'(t)$	$b_{32} = (S2X)/84$
		$b_{31} = [(S1X) -196b_{33}$
		$b_{30} = -4b_{32}$

Smoothed values $x'(i)$ can then be calculated for any i using the selected order polynomial.

E. SPECIAL TREATMENT (STEP 5)

There are three conflicting properties desired in a data smoothing algorithm. One of these is the extraction of the maximum information from the data on the vehicular path. The second is that automation of the algorithm (the computer program) be as complete as possible so that little or no subsequent manual processing is required. The third property is that the computer program to implement the algorithm be as simple as possible.

If the first property were omitted or if no outliers or missing points were present in the data then STEP 5 could be deleted. The purpose of STEP 5 is to extract more information on the vehicular path from data in the vicinity of multiple outliers and/or missing points. This will be more difficult to automate (awkward to program) and the resulting information will be somewhat degraded in quality. Nevertheless, efforts to implement this step are important since path segments with multiple outliers and/or missing points appear to occur when information on torpedo and target locations are most important (e.g., in the vicinity of intercept).

There are several possibilities for using least-squares smoothing at data points in a data segment other than the midpoint and even for estimation or prediction at points outside the segment. These possibilities have not been fully explored or developed to the state where they can be specified for inclusion in STEP 5.

F. OTHER FEATURES OF LEAST-SQUARES SMOOTHING

1. Appropriate Polynomial Degree

As discussed in Reference 1 and in Section I of this report each observed datum is assumed to contain two components, one representing the actual coordinate of a point on the path of a moving vehicle and the other a noise component resulting from inaccuracy of measuring the first. Thus, for the observation x_i at time t_i, we have

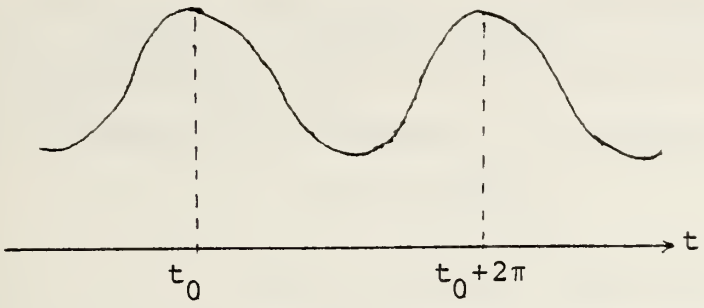
$$x_i = P(t_i) + n_i \ .$$

If the data segment to be fitted is short, it will be assumed that the path component $P(t_i)$ for any of the coordinates can be represented by a low order polynomial with its error in fitting the actual path component being small in comparison to the noise component.

The highest order polynomial required for a given length of data segment depends upon the turning rate of the vehicle being tracked and the data tracking rate. For example, a torpedo is capable of making a complete circle within 10 data intervals (data segments of 11 points). The x-component for a circular path can be represented as a sine function, e.g.,

$$P_x(t) = a + b \sin(c+\sigma t)$$

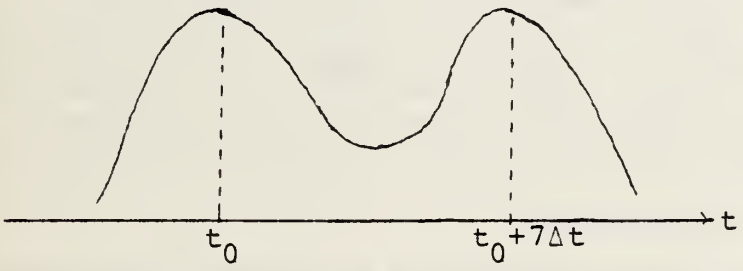
which might be graphed as below.



The minimum order polynomial which might be considered as a reasonable representation of this is four so that

$$P(t) \doteq b_0 + b_1 t + b_2 t^2 + b_3 t^3 + b_4 t^4$$

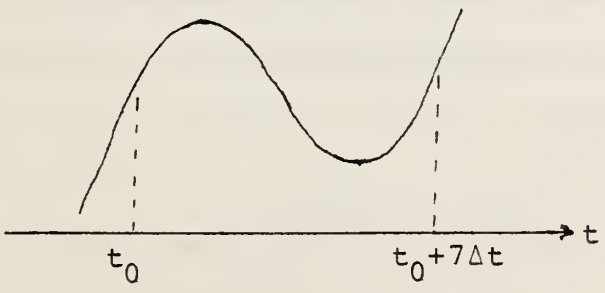
with a graph of the form shown below,



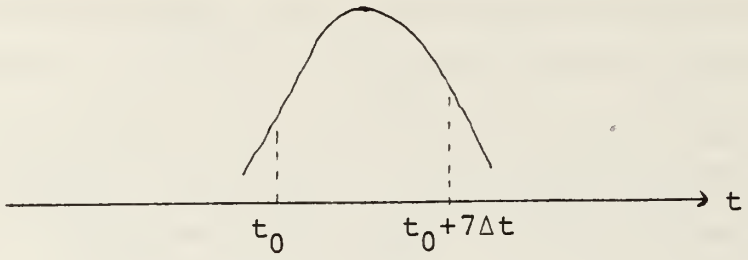
For data segments of 7 points, it would appear that a polynomial of order three (cubic) might be adequate so that

$$P(t) \doteq b_0 + b_1 t + b_2 t^2 + b_3 t^3$$

with a graph below.



For a vehicle maneuvering at a lower rate a polynomial of second order (quadratic or parabolic) or first order (linear) might be adequate. The graph of a parabola could appear as below.



Examination of the results of fitting 7-point data segments of the x-component of a torpedo run at NUWES which were presented in Table 1 illustrate the capabilities of polynomials of orders three or less to represent actual path components.

The interaction of data segment length and corresponding polynomial order requirements has not been explored.

2. Quality of Smoothing

The "quality" of fit of a polynomial to a data segment would appear to be better represented by the quantity $\sum e_i^2$ than the sample standard deviation SEK which was used in Section III D do select the appropriate order polynomial to fit a data segment.

There are two aspects of the use of SEK as a measure of the quality of smoothing that need further examination. One of these is the effect of missing and/or outlier points on the formula for SEK. Thus

$$SEK = \sqrt{(\sum e_{Ki}^2/r}$$

where the "degrees of freedom" divisor, μ, for a data segment of n observed values is

$$r = n - (K+1) .$$

When there are m temporary values resulting from missing and/or outlier data values in the segment a more appropriate formula is

$$r = (n-m) - (K+1).$$

The presence of temporary values in a data segment decreases r producing an increase in SEK indicating a larger noise component and hence a lower quality of the smoothed values. Thus SEK can indicate the reduction in the "quality" of smoothing by the presence of temporary values in the data segment. In the example where $n = 7$, $K = 3$ and $m = 3$, we have $r = 0$. A third order polynomial ($K=3$) can be fitted exactly to the remaining $n - m = 4$ observed data values including their noise component and no smoothing has been performed. Also, no estimate of the noise component (σ_N) is possible.

Increasing the length of data segments (n) could be used to increase r provided higher order polynomials (increasing K) is not also required. The possibility of r increasing when n is increased should also be considered.

The other aspect of the use of SEK as a measure of the "quality" of smoothed values is the degredation in the quality depending on the location of the smoothed value with respect to the center of the data segment used for smoothing. This can be

demonstrated by the formula for a confidence interval for the actual value $\tilde{X}(i)$ when the actual relationship is $\tilde{x}(t) = a_0 + a_1 t$ and a polynomial of order one ($x'(i) = b_0 + b_1 t$) is fitted by least squares to a 7-point data segment. The confidence interval for $\tilde{X}(i)$ is of the form

$$\left(x'(i) - c\sqrt{\frac{1}{7} + \frac{i^2}{28}}\ SE1\ ,\ x'(i) + c\sqrt{\frac{1}{7} + \frac{i^2}{28}}\ SE1 \right)$$

The factor $\sqrt{\frac{1}{7} + \frac{i^2}{28}}$ produces an increase in the width of the confidence interval with the distance $|i|$ from the center of the 7-point data segment. This can be considered as a decrease in the "quality" of $x'(i)$ as an estimate of $\tilde{X}(i)$. This factor could also be combined with SE1 so that $\sqrt{\frac{1}{7} + \frac{i^2}{28}}\ SE1$ could be treated as the appropriate standard deviation value for the noise component at time i. This apparent increase in the standard deviation of the noise also represents a decrease in the quality of the smoothed value at time i.

A general form for confidence intervals about a second order polynomial can be obtained using material in the Appendix B-3 of Reference 1, Notations are differences in Reference 1 and translation will be required to make it applicable to the material in this report. A corresponding confidence interval for third order polynomials needs development.

The requirement that 7-point data segments for smoothing the value at time t_i also be centered at time t_i was specified by the information that confidence intervals have

minimal confidence interval width at this point in the sample when the fitted polynomial is linear. There is some evidence (Ref. 1) that the minimum width of confidence intervals does not occur at the sample midpoint when the fitted polynomial is quadratic.

IV. SUMMARY AND RECOMMENDATIONS

The Algorithm presented in Section II and discussed in Section III provides a reasonable approach for use in smoothing 3-D data at NUWES. In the proposed algorithm, the data is first screened for outliers (wild data) using fourth order sequential differences and then a special form of least-squares smoothing is performed to fit a low-order polynomial to data segments.

The presence of multiple outliers and/or missing values has two major effects on data smoothing.

(a) They make smoothing of observational data difficult in thei vicinity. (Only preliminary treatment of this problem is included in the algorithm.)

(b) They degrade the quality of smoothing in their vicinity. (This is discussed but not formally incorporated into the algorithm.)

Although the algorithm can be used in its present form, improvements can, and should, be considered. Some directions for improvements are:

(a) The data segment length (7 points) and the degree (3 or less) for fitting polynomials was somewhat arbitrarily selected. The following possible changes need examination:

  (i) Increase the data segment length from 7 to 9 and fitting polynomials of degree 3 or less.

  (ii) Increase the data segment length to 11 and fitting polynomials of degree 4 or less.

(b) Special treatment of multiple outliers and/or missing points needs additional effort to achieve the goal of full automation.

(c) Development, and inclusion in the algorithm, of some measure of the quality of smoothing could be of considerable interest to users of the smoothed data. It would be useful even when no outliers and/or missing points are present in the data and its usefulness increases as an indicator to potential users of the extent of degradation when the data includes outliers and/or missing points.

REFERENCES

1. J. B. Tysver, "Smoothing 3-D data for torpedo paths", Naval Postgraduate School Technical Report NPS55-78-036Pr, May 1978.

2. J. B. Tysver, "Use of sequential differences in smoothing 3-D data", Naval Postgraduate School Technical Report NPS55-79-012Pr, May 1979.

3. H. J. Larson, Introduction to Probability Theory and Statistical Inference, Second Edition, John Wiley and Sons, 1974.

DISTRIBUTION LIST

	No. of Copies
Commanding Officer Attn: Mr. R. L. Marimon, Code 70 Naval Undersea Warfare Engineering Station Keyport, WA 98345	2
Library, Code 0142 Naval Postgraduate School Monterey, CA 93940	1
Dean of Research Code 012A Naval Postgraduate School Monterey, CA 93940	1
Professor J. B. Tysver Code 55Ty Naval Postgraduate School Monterey, CA 93940	10
Naval Undersea Warfare Engineering Station Keyport, WA 98345	
Attn: Code 51	1
Code 52	1
Code 53	1
Code 54	1
Code 5122	2
Code 50 Attn: R. Mash	1
Code 80 Attn: CDR C. Gertner	2
Code 0115-S General Administration	1
Code 0116, Technical File Branch	1
Naval Postgraduate School Monterey, CA 93940	
Attn: Prof. O. B. Wilson, Code 61Wl	3
Prof. H. A. Titus, Code 62Ts	1

DISTRIBUTION LIST

		No. of Copies
Naval Postgraduate School Monterey, CA 93940		
Attn:	Code 55Mt	1
	Code 55As	1
	Code 55Bn	1
	Code 55Bw	1
	Code 55Cu	1
	Code 55Ei	1
	Code 55Ey	1
	Code 55Fo	1
	Code 55Gv	1
	Code 55Hh	1
	Code 55Hk	1
	Code 55Hl	1
	Code 55Jc	1
	Code 55La	1
	Code 55Lw	1
	Code 55Ls	1
	Code 55Mg	1
	Code 55Mh	1
	Code 55Mu	1
	Code 55Mp	1
	Code 55Ni	1
	Code 55Py	1
	Code 55Pk	1
	Code 55Re	1
	Code 55Rh	1
	Code 55Ro	1
	Code 55Sy	1
	Code 55Su	1
	Code 55Ta	1
	Code 55Tw	1
	Code 55Ty	1
	Code 55Ws	1
	Code 55Ze	1